

2006

# Data Distribution and Archiving in Support of the Agricultural Ecosystems Program

Gail Steinhart  
*Cornell University*

Follow this and additional works at: <https://surface.syr.edu/nyscilib>

 Part of the [Library and Information Science Commons](#)

---

## Recommended Citation

Steinhart, Gail, "Data Distribution and Archiving in Support of the Agricultural Ecosystems Program" (2006). *Upstate New York Science Librarians Conference*. 60.  
<https://surface.syr.edu/nyscilib/60>

This Presentation is brought to you for free and open access by SURFACE. It has been accepted for inclusion in Upstate New York Science Librarians Conference by an authorized administrator of SURFACE. For more information, please contact [surface@syr.edu](mailto:surface@syr.edu).



Cornell University  
Albert R. Mann Library

# Data Distribution and Archiving in Support of the Agricultural Ecosystems Program

Gail Steinhart  
Research Data &  
Environmental Sciences Librarian  
Albert R. Mann Library  
Cornell University

*Upstate Science Librarians, October 20, 2006*



# Collaborators

## Cornell departments and units:

- Animal Science
- Biological and Environmental Engineering
- Crop and Soil Science
- Ecology and Evolutionary Biology
- Horticulture
- Natural Resources
- Mann Library

## Other organizations:

- Cornell Cooperative Extension of Chemung County
- Institute of Ecosystem Studies
- Univ. Maryland Center for Environmental Science
- Univ. Nebraska-Lincoln School of Natural Resources
- Upper Susquehanna Coalition

Funding: USDA Cooperative State Research, Education,  
and Extension Service



# Research context

- The Susquehanna River watershed contributes a large portion of the nutrients and sediments impacting the Chesapeake Bay, which suffers from excessive nutrient and sediment inputs.
- Improving nutrient and erosion controls in the watershed of the Susquehanna River, the largest river entering the Bay, is one way to improve the health of the Bay itself. New York is committed to reduce the impact of its part of the Susquehanna River watershed on the Bay.
- The research project is designed to better understand the sources and sinks of nutrients and sediments in the New York portion of the Susquehanna watershed, using this as a model for rural landscapes in general.



# Goal

*The AEP Upper Susquehanna River Basin project should serve as a source of valuable data and insight for people in county, state, and the Federal government, NGOs, and the public to improve water quality in the Upper Susquehanna Basin and the Chesapeake Bay.*

**Specific Objective:** Provide an easy way for the public to learn about the scientific goals and results of the AEP Upper Susquehanna River Basin project.

## **Approach:**

- Document data sets using Ecological **Metadata** Language (EML).
- Deposit and preserve data sets in Cornell's digital repository, **DSpace**.
- Create a public research **portal**, with background information, research plans, results, and data.



# Goal

*The AEP Upper Susquehanna River Basin project should serve as a catalyst for innovative cross-disciplinary research at Cornell that will produce better scientific understanding of nitrogen, phosphorus, and sediment cycling in the Upper Susquehanna basin and the Chesapeake Bay.*

**Specific Objective:** Provide an easy way for researchers to know what other project participants have already done, are doing, and are planning to do (who, what, when, where, why).

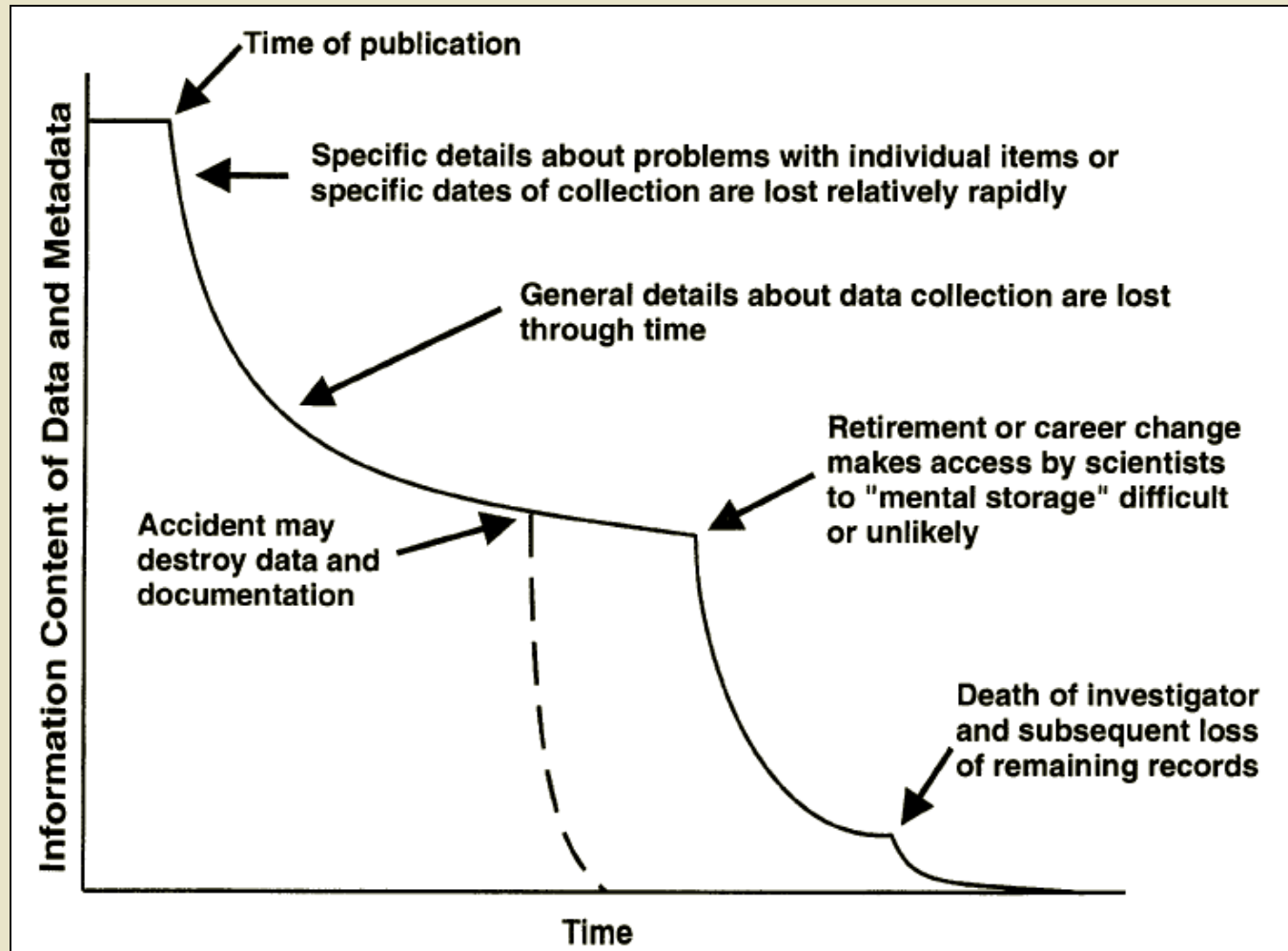
**Approach:** Create a **wiki**, with access restricted to project participants. The wiki can be used to share datasets, preliminary results, and any other information about the project as it progresses, before such information is made publicly available.



# Metadata

- Metadata serve as documentation for data, describing the content, purpose, structure, format, and accessibility of datasets. Interdisciplinary and collaborative science creates an important demand for a set of “instructions” for researchers to make sensible judgments about whether and how they might use data provided by their colleagues.
- Metadata also serve a functional role in digital repositories, providing the raw material that makes it possible to display information about a dataset, and for search engines to index repositories and deliver results to users.

# Avoid data entropy!



Michener et al., 1997





# Ecological Metadata Language: EML

- Developed specifically for ecological data (ESA, LTER)
- Modular and extensible XML-based standard
- Accommodates information on methods, geographic coverage, temporal coverage, detailed descriptions of tabular data
- <http://knb.ecoinformatics.org/software/eml/>
- Comes with tools!

# Morpho

- Easy to use, platform independent metadata editor.
- Allows users to upload metadata and data to a server.
- Allows users to search, view, and save public data and metadata. (*Interacts with Metacat*)

**New Data Package Wizard**  
Welcome to the New Data Package Wizard

This wizard creates a *Data Package*, consisting of the structured documentation that describes your data (i.e., metadata), and the data themselves.

If you wish to improve your understanding of metadata and related concepts, you can read [An Introduction to Ecological Metadata Language \(EML\)](#), which provides background information and examples of metadata. The wizard uses EML, and you may need to adequately document your data, use *Morpho Editor* (after you finish this wizard) on the main Morpho screen).

Before beginning you should have your data (electronic or hardcopy format) available.

- ◆ **Title and abstract**
- ◆ **Keywords**
- ◆ **People and Organizations**
- ◆ **Usage Rights**
- ◆ **Research Project Information**
- ◆ **Coverage Details**
- ◆ **Methods and Sampling**
- ◆ **Access Information**

**Note:** Required information includes the title and personnel information for your data. It is highly recommended that you fill in as much as possible.

Step 1 of 15

---

**New Data Package Wizard**  
**Title and Abstract**

**Enter the title of the data package.** The title field provides a description of the data that is long enough to differentiate it from other similar data. e.g. Vernal Pool Amphibian Density Data, Isla Vista, CA USA, 1990-1996

Title:

**Enter an abstract that describes the data package.** This abstract is used to describe the data package. You may want to describe the objectives, key aspects, design, etc.

Abstract:

Step 2 of 15

---

**New Data Package Wizard**  
**Define Attribute or Column:**

**Name:**  Name of the attribute as it appears in the data file

**Label:**  A more readable label for the attribute

**Definition:**  Define the contents of the attribute (or column) precisely, so that a data user could interpret the attribute accurately.  
e.g. "spdm" is the number of individuals of all macro invertebrate species found in the plot

**Storage:**  Storage type for this field e.g. integer, float

**Storage System:**  The system used to define the storage types e.g. C, Java, Oracle

Unordered: unordered categories or text (statistically **nominal**) e.g. Male, Female

Ordered: ordered categories (statistically **ordinal**) e.g. Low, High

Relative: values from a scale with equidistant points (statistically **interval**) e.g. 12.2 meters

Absolute: measurement scale with a meaningful zero point (statistically **ratio**) e.g. 273 Kelvin

Date-Time: date or time values from the Gregorian calendar e.g. 2002-10-24

**Relative**

**Standard Unit:**

**Precision:**  e.g. for an attribute with unit "meter", a precision of "0.1" would be interpreted as precise to the nearest 1/10th of a meter

**Number Type:**

**Bounds:**

# EML record

```

--<eml:eml packageId="gss1.15.2" system="knb" xsi:schemaLocation="eml://ecoinformatics.org/eml-2.0.1 eml.xsd">
- <dataset>
- <title>
  Lake Ontario Embayments - temperature and dissolved oxygen profiles
</title>
- <creator id="1144096749380">
- <individualName>
  <givenName>Gail</givenName>
  <surName>Steinhart</surName>
</individualName>
<organizationName>Center for the Environment</organizationName>
<positionName>Research coordinator</positionName>
- <address>
  <deliveryPoint>Rice Hall</deliveryPoint>
  <deliveryPoint>Cornell University</deliveryPoint>
  <city>Ithaca</city>
  <administrativeArea>NY</administrativeArea>
  <postalCode>14853</postalCode>
</address>
</creator>
- <abstract>
- <para>
  Temperature and dissolved oxygen profiles collected as part of and NSF funded project (O
  Biocomplexity. Physical, biological, and human interactions shaping the ecosystems of freshw
</para>
</abstract>
- <keywordSet>
<keyword>Lake Ontario</keyword>
<keyword>Blind Sodus Bay</keyword>
<keyword>Little Sodus Bay</keyword>
<keyword>Sterling Pond</keyword>
<keyword>Juniper Pond</keyword>
<keyword>South Sandy Pond</keyword>
<keyword>North Sandy Pond</keyword>
<keyword>Colwell Pond</keyword>
<keyword>Floodwood Pond</keyword>
<keyword>limnology</keyword>
</keywordSet>
- <intellectualRights>
- <para>
  Protected Data Data is freely shared within the research group. However, findings or conclu
  another individual's data should be brought to that individual's attention. Only the owner of a
  individuals not affiliated with the research group. Acknowledgement of Support and Disclaim
Done

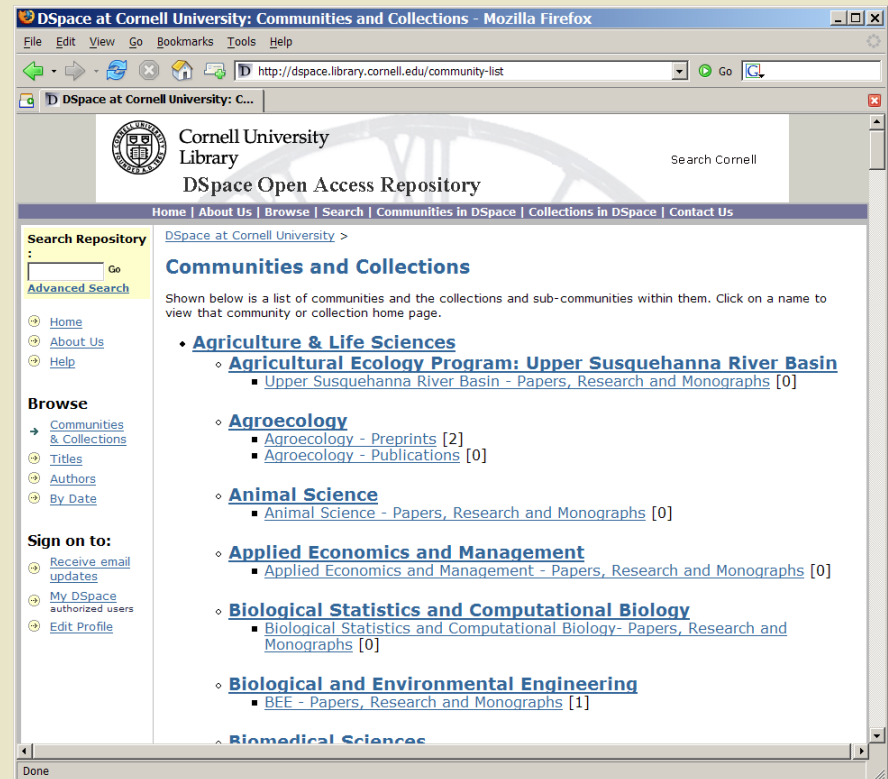
```

The screenshot shows a web browser window titled "Data Package: gss1.15.2". The page content includes the following information:

- Title:** Lake Ontario Embayments - temperature and dissolved oxygen profiles
- Accession Number:** gss1.15.2
- Keywords:** Lake Ontario, Blind Sodus Bay, Little Sodus Bay, Sterling Pond, Juniper Pond, South Sandy Pond, North Sandy Pond, Colwell Pond, Floodwood Pond, limnology.
- Data Set Owner(s):**
  - Individual: Gail Steinhart
  - Organization: Center for the Environment
  - Position: Research coordinator
  - Address: Rice Hall, Cornell University, Ithaca, NY 14853
- Abstract:** Temperature and dissolved oxygen profiles collected as part of and NSF funded project (OCE-0083625): Biocomplexity: Physical, biological, and human interactions shaping the ecosystems of freshwater bays and lagoons.
- License and Usage Rights:** Protected Data Data is freely shared within the research group. However, findings or conclusions made while using another individual's data should be brought to that individual's attention. Only the owner of a dataset may share it with individuals not affiliated with the research group. Acknowledgement of Support and Disclaimer This research was supported by Biocomplexity award number OCE-0083625 from the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
- Geographic Coverage:** Selected Lake Ontario embayments in New York state: Blind Sodus Bay, Little Sodus Bay, Sterling Pond, Juniper Pond, South Sandy Pond, North Sandy Pond, Colwell Pond, Floodwood Pond
- Bounding Coordinates:**
  - West: -77.25 degrees
  - East: -76.125 degrees
  - North: 44.0 degrees
  - South: 43.25 degrees

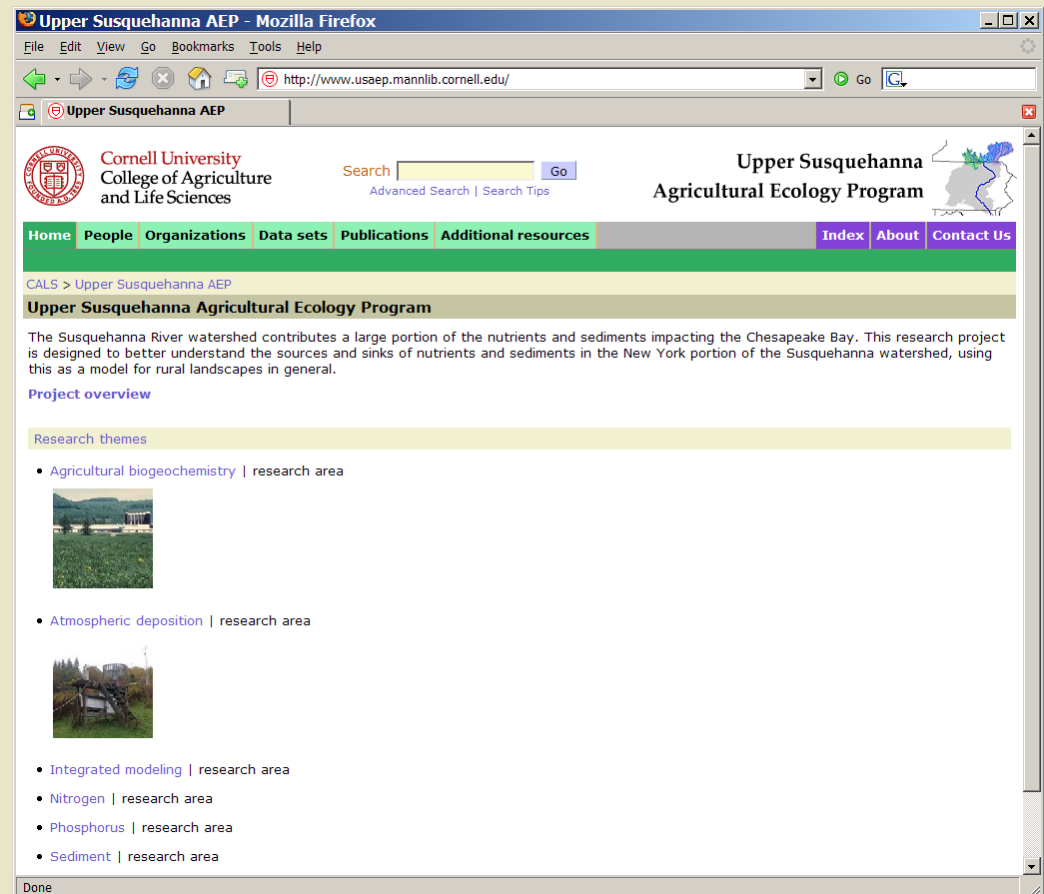
# DSpace

- Open source
- Easy submission
- Indexable by search engines
- Communities, collections
- Customizable workflow
- Supports OAI-PMH
- *Already implemented at Cornell*



# Portal

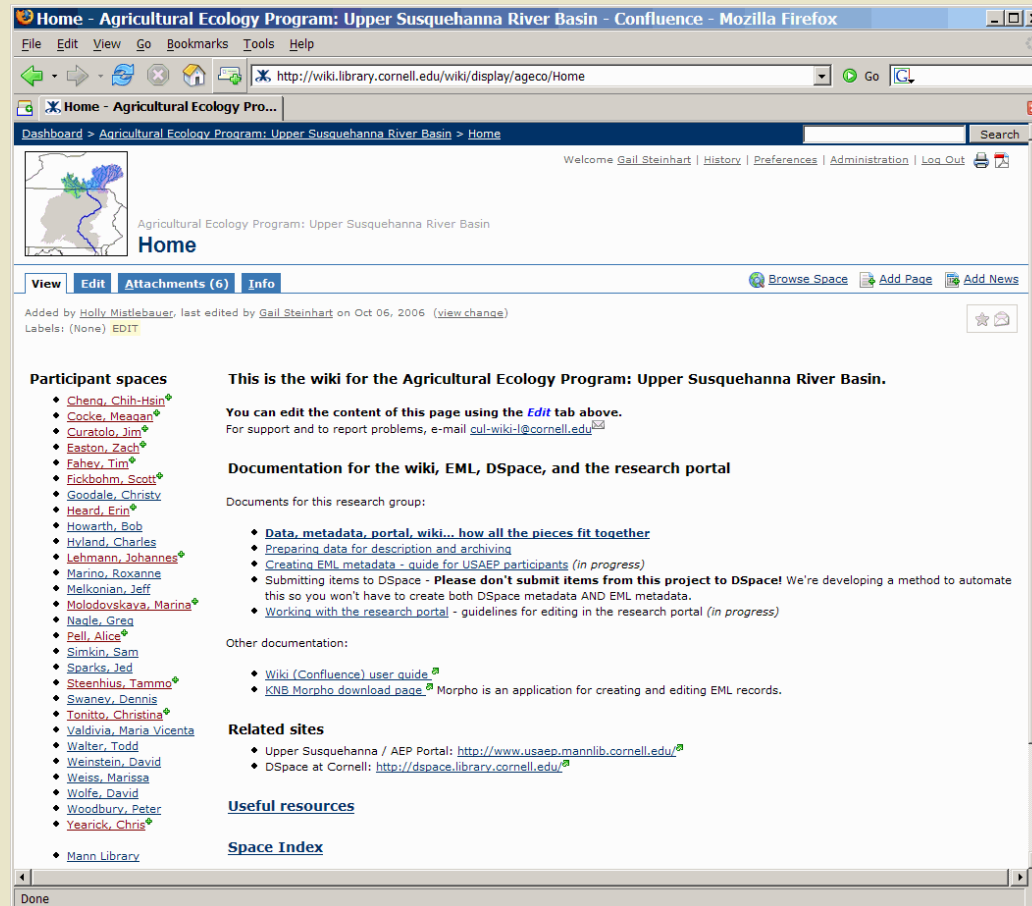
- <http://www.usaep.mannlib.cornell.edu/>
- Modeled after Vivo and CALS research portals
- Presents information on the project *in context*
- Links to data in DSpace



The screenshot shows a Mozilla Firefox browser window displaying the website for the Upper Susquehanna Agricultural Ecology Program. The browser's address bar shows the URL <http://www.usaep.mannlib.cornell.edu/>. The website header includes the Cornell University logo and the text "Cornell University College of Agriculture and Life Sciences" on the left, and "Upper Susquehanna Agricultural Ecology Program" on the right. A search bar is located between the two. Below the header is a navigation menu with links for Home, People, Organizations, Data sets, Publications, Additional resources, Index, About, and Contact Us. The main content area features a section titled "Upper Susquehanna Agricultural Ecology Program" with a brief description of the project's focus on the Susquehanna River watershed. Below this is a "Project overview" section and a "Research themes" section listing several research areas: Agricultural biogeochemistry, Atmospheric deposition, Integrated modeling, Nitrogen, Phosphorus, and Sediment. Each research area is accompanied by a small thumbnail image.

# Wiki

- Share information within the group
- Gaining traction in the sciences:
  - <http://www.openwetware.org/>
  - [“Science in the Web Age: Joint efforts”](#)



The screenshot shows a Mozilla Firefox browser window with the address bar displaying <http://wiki.library.cornell.edu/wiki/display/ageco/Home>. The page title is "Home - Agricultural Ecology Program: Upper Susquehanna River Basin - Confluence - Mozilla Firefox". The page content includes a map of the Upper Susquehanna River Basin, a "Home" heading, and several sections:

- Participant spaces:** A list of names with small icons, including Cheng, Chih-Hsin; Cocke, Meagan; Curatolo, Jim; Easton, Zach; Fahey, Tim; Fickbohm, Scott; Goodale, Christy; Heard, Erin; Howarth, Bob; Hyland, Charles; Lehmann, Johannes; Marino, Roxanne; Melkonian, Jeff; Molodovskaya, Marina; Nagle, Greg; Pell, Alice; Simkin, Sam; Sparks, Jed; Steenhius, Tammo; Swaney, Dennis; Tonitto, Christina; Valdivia, Maria Vicenta; Walker, Todd; Weinstein, David; Weiss, Marissa; Wolfe, David; Woodbury, Peter; Yearick, Chris; and Mann Library.
- This is the wiki for the Agricultural Ecology Program: Upper Susquehanna River Basin.** A section stating "You can edit the content of this page using the [Edit](#) tab above." and providing an email address for support: [cul-wiki-l@cornell.edu](mailto:cul-wiki-l@cornell.edu).
- Documentation for the wiki, EML, DSpace, and the research portal:** A section titled "Documents for this research group:" listing items like "Data, metadata, portal, wiki... how all the pieces fit together", "Preparing data for description and archiving", "Creating EML metadata - guide for USARP participants (in progress)", "Submitting items to DSpace - Please don't submit items from this project to DSpace!", and "Working with the research portal - guidelines for editing in the research portal (in progress)".
- Other documentation:** A section listing "Wiki (Confluence) user guide" and "KNB Morpho download page" with a note that Morpho is an application for creating and editing EML records.
- Related sites:** A section listing "Upper Susquehanna / AEP Portal: <http://www.usaep.mannlib.cornell.edu>" and "DSpace at Cornell: <http://dspace.library.cornell.edu>".
- Useful resources** and **Space Index** sections.

Upstate Science Librarians, October 20, 2006



# Putting it all together

- Project meetings
- Metadata / Morpho workshop
- Documentation
- Individual consulting



## Early findings...

- USDA proposal
- Scope creep
  - Data formats
  - Files sizes
  - Need to share large files, privately
- Portal maintenance
- Extensible model?





*Thank you*

Gail Steinhart  
Research Data & Environmental Sciences Librarian  
Albert R. Mann Library, Cornell University

GSS1@cornell.edu